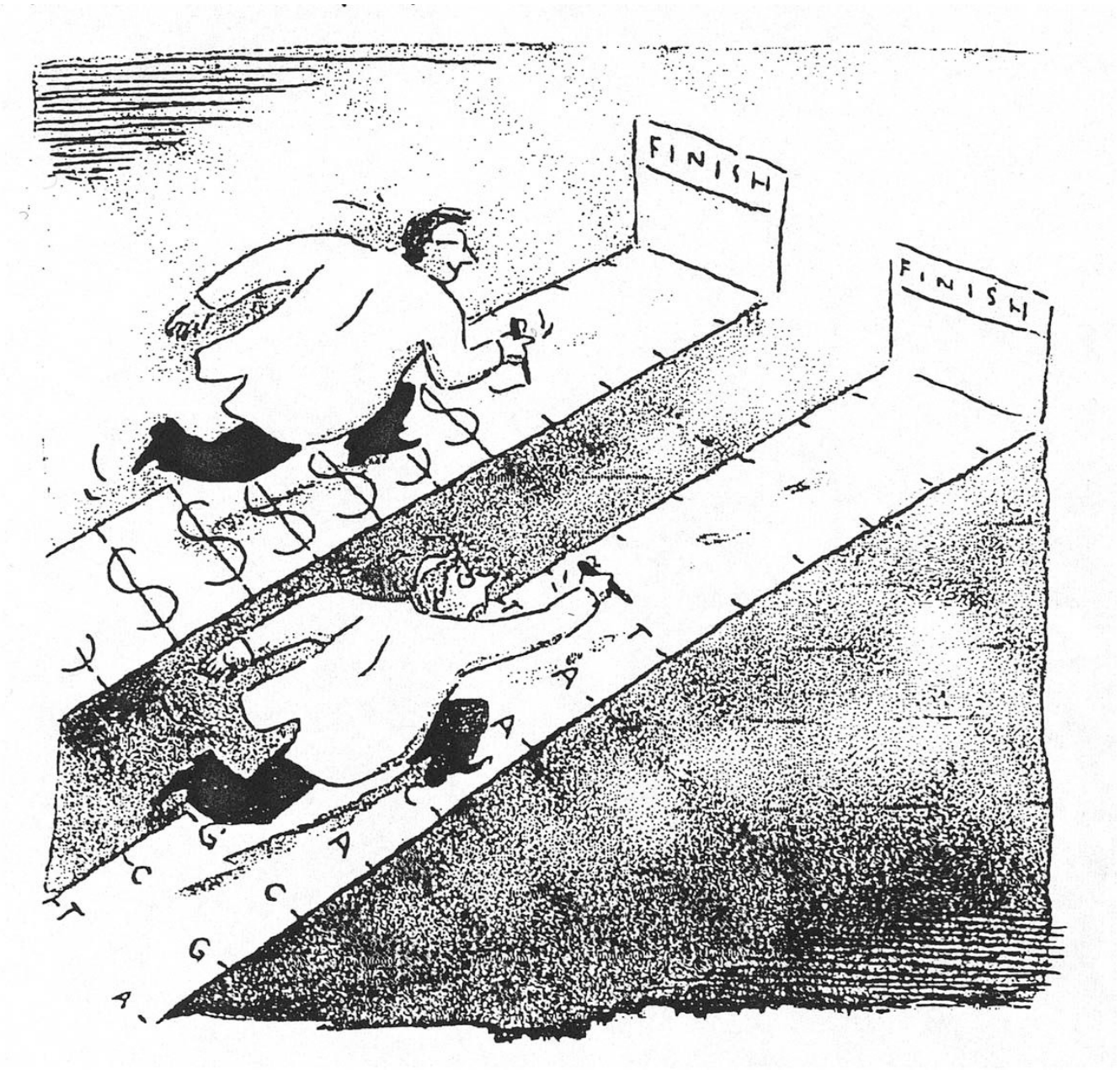
Open science, Open data, Open products

Tim Hubbard
Sanger Institute
timjph@gmail.org

Human genome race

won by public project

open access for all



International agreement on data release

"All human genomic sequence information should be freely available and in the public domain in order to encourage research and development and to maximise its benefit to society."

The Bermuda Statement, February 1996

Assemblies of 1-2 kb are deposited in public database (GenBank, EBI) every 24 hours

No patents are filed

Evolution of Data Release Policy

- Bermuda principles reaffirmed at January 2003 NIH/WT meeting*, leading to new NIH/WT policy to divide funding into two classes:
  - R01 projects:
    - Competitive
    - Release data on publication
  - "Community Projects"
    - Non-competitive
    - Managed
    - Release data in real time

*Nature 421 , 875 (2003)

Open data partnerships

- Example:
  The SNPs consortium (TSC):
  - 12 companies & Wellcome Trust
  - $3 million membership
  - >1 million SNPs data freely accessible, August 2000

- Why?
  - The more people analysing a block of data, the more valuable it is.

- Internet context of open data:
  - create “level playing field”
  - Aids capacity building

Problems to solve

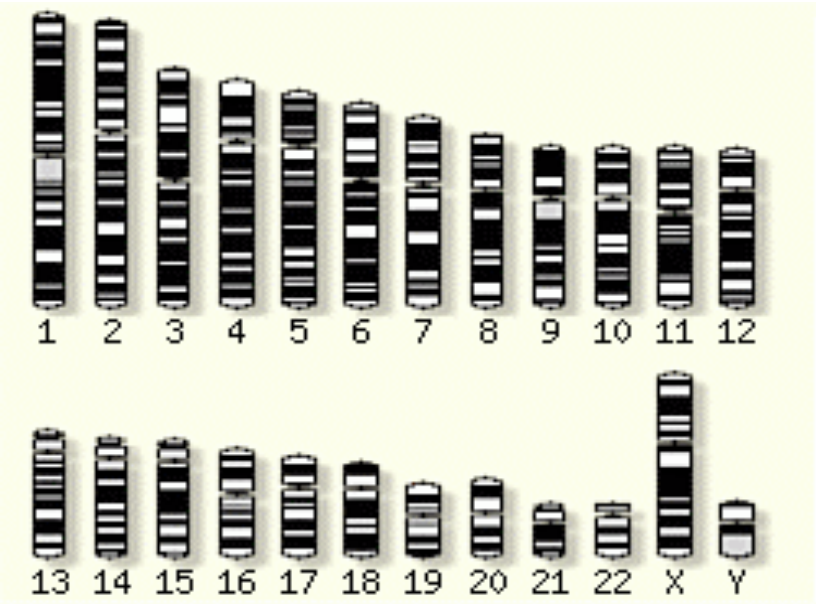
- Cultural attitudes towards data sharing
  - New ways of allocating credit
  - Adjustment to more competitive environment
- Practical issues of data sharing
  - Standardization of data sets
  - Engineering to allow distributed data access
  - Stable infrastructure funding to support data archives

Rufus' open knowledge principles:

- Incremental
- Decentralised
- Collaborative
- Componentised

Coordinate Systems

- Natural coordinate system: chrX:1-n



- Genomes sequenced as thousands/millions of fragments
  - uncertain order
  - errors
  - Gaps
- Many transformations between coordinate systems
  - Simple: genome to genome
  - Exotic: genome to protein

Genome resources deliver more than sequence

```
31861 ggaaaaattaagttttagaagtgtttaaggtacttttttctataattatttattataaaag
31921 atatagtcttcccttgtcatgacatgtggttaattctatgaaagtttgatagaattatga
31981 tattcacataaaacaaggttgtgatgtctggtgtttcacagtccgttgacttttgatgca
32041 attcttggtagacatcctccgactatgttttagatgtcattttcaagttttgcagtttct
32101 cgaaatattagaagccatgtctgcaccgaactgcgcacgaaaatatgatattgctcgtct
32161 ttccagcttgaattttcaaatttcccaatatgtttatcttagcttgataagcttaacttt
32221 tatattttcttattttgctgtgaaaattgttcatcaaaaatcgattttccaactttccac
32281 taaaatcttattatttcacaatttggtttctgcgaatcttcatcaacttttatacttatt
32341 ttccgcactccgaaggctcaacctggcatatttctatattgacgaaccatgtgttcctct
32401 aatatcggaagctgactgtctaccttacctcaaggttttagtaactggaataagtggaat
32461 gatctatggccaaactggtctacttttggaacgaggttgtgcaactttcatcaaagatta
32521 tgataagaagacaagcatgtttgttgggatagcaatatcaattgccattttgtttttgag
32581 cttgatcactggaaaaattataatttgggatgacccacttcaaggatatcttctttcgtg
32641 tgtttcctatccaagtcaaagtgttgaaagatcaagactatttgcaagtatttatacatt
32701 tatatcttcattcaatttggtttctcagttttgttaagaagatataacaaaaaactgga
32761 atattcgtaagctttctaagttgtttgaaaagtttactgttcgtcaatttctggaatttt
32821 agttaagaaagtcagatggcaaatcatgatgccttcataaaatgagtaaataaacctgat
32881 tagtttactattttgttcaaacttcaattttggaacgtattgcgcttaaaggtactgaaa
32941 actagtatgcacgaaaaacttcttactgtctactagatatctttaattgcctgaaacgcg
33001 gcaatatttagtgcaattcaacttccagacgtttgactctttgtaatttacttttgcgta
33061 atatctgatctctgaaatttctgaatagtatttttctgattagcttgtttttcttctcat
33121 tgtttccactacatttgcttccaaacttggaaaacaaatttttgaataaatctagaatat
33181 tctaatctggtttttgatgttttaaagttccattaatgtttttttgagcgtaagaaatgtt
33241 tcaattttccagaacacccttttgttgttggcccgagatttcggaaaagagaagtgattga
33301 ctccacaagtacaatatgctttttgacatttgttcaattcatattcatgttcatttattc
33361 attcggaatattcacactgaaaactattcgaagcatgcttacttacagacagtactattt
33421 cattgttgtctggttctatgtaagcacttgaaatttatttttaaaaagctgaaaattttat
33481 ttccagacaattccattcattgctgcgctgtttccaattctactagtttacaggattcgt
33541 tcttcacatgttagccgagtaacgattattaaaacatttacaaaaaccaagcaaacacaa
33601 gaagaacacattaagcaattgaaaaacgtttggaattgaacataatgattcctattttat
33661 gaaatctgaatttttgtaaatatgtgtatatttttttggaataaataattgtcattagga
33721 aaaaaatcgagtgatcttcttttccgaattttcattttaatttcgagatagtaagaaaag
33781 ttgcaagtcatttgaattcaacgattttccttaatatttctgaatttattcttcaaatgt
33841 at
```

Less than 1/1,000,000 part of the genome sequence

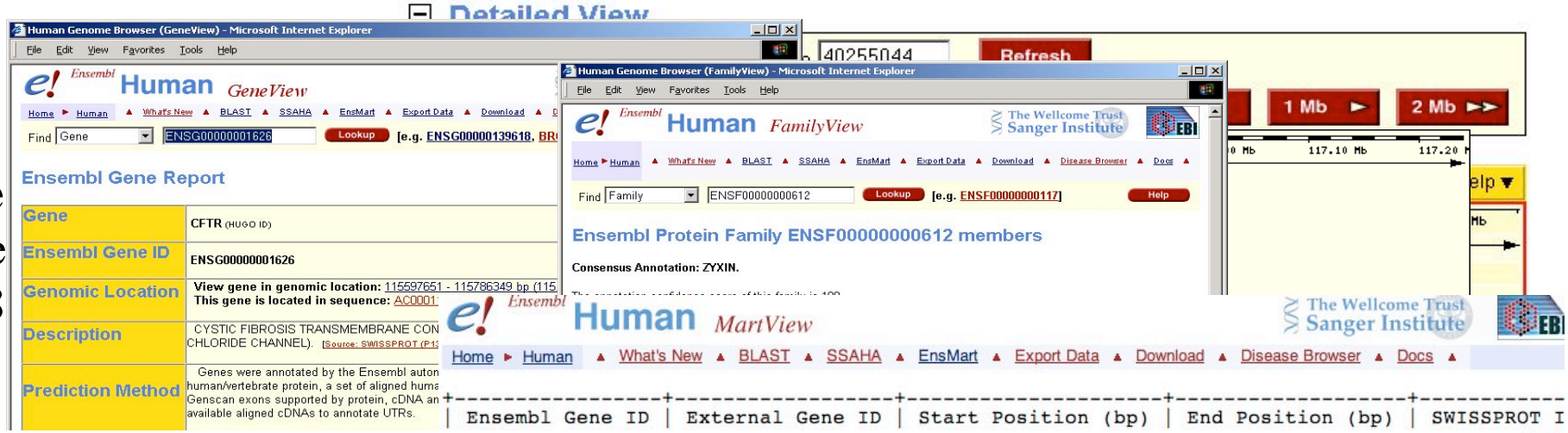
project Ensembl

wellcome trust sanger institute

EBI

Gene
(Ense
22,28

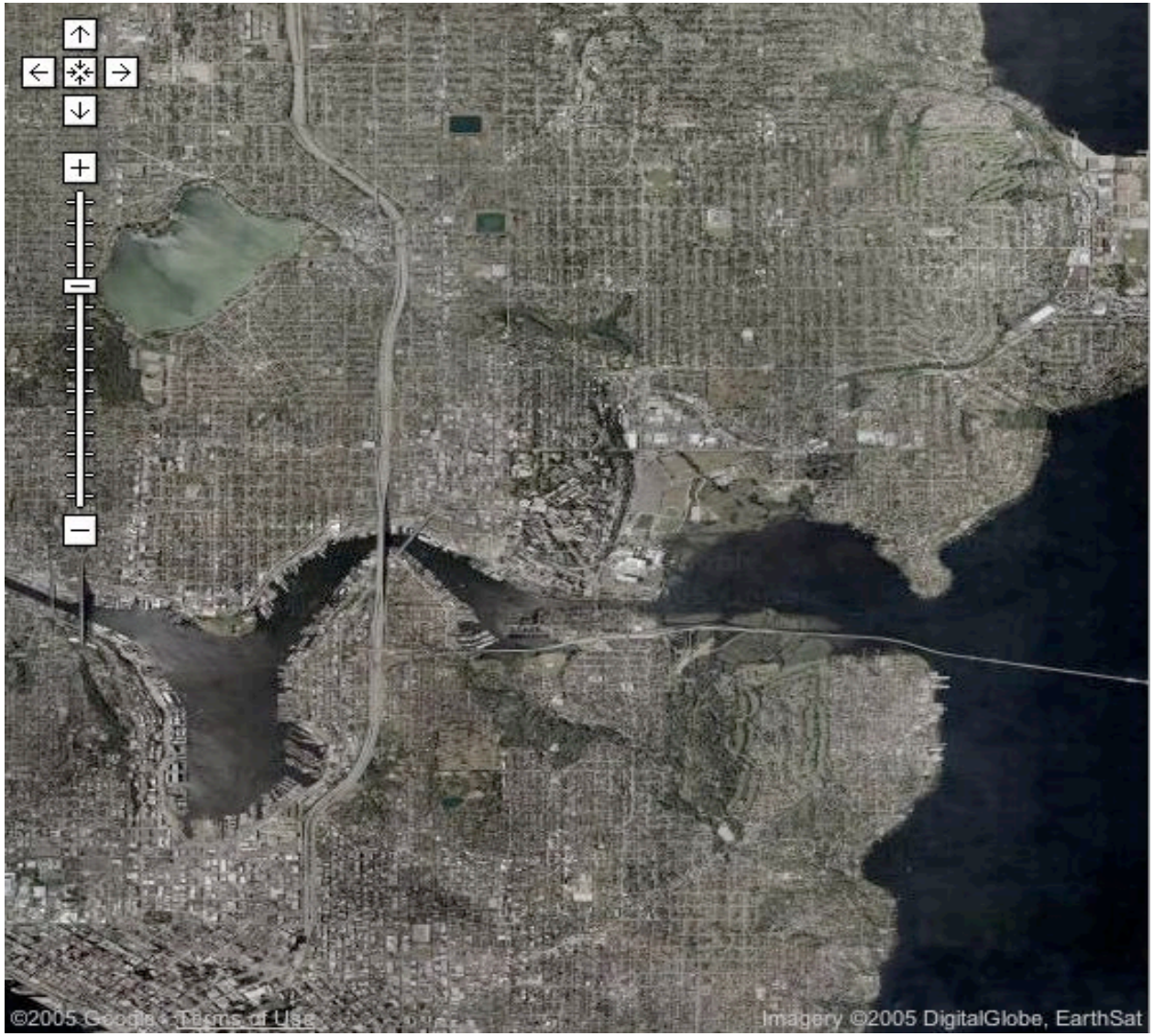
Other
~80 di
of info


Human Genome Browser (GeneView) - Microsoft Internet Explorer
Ensembl Human GeneView
Find Gene ENSG00000001626 Lookup [e.g. ENSG00000139618, BR
Ensembl Gene Report
Gene CFTR (HUGO ID)
Ensembl Gene ID ENSG00000001626
Genomic Location View gene in genomic location: 115597651 - 115786349 bp (115
This gene is located in sequence: AC0001
Description CYSTIC FIBROSIS TRANSMEMBRANE CON
CHLORIDE CHANNEL). [Source: SWISSPROT (P1
Prediction Method Genes were annotated by the Ensembl auton
human/vertebrate protein, a set of aligned huma
Genscan exons supported by protein, cDNA an
available aligned cDNAs to annotate UTRs.
Human Genome Browser (FamilyView) - Microsoft Internet Explorer
Ensembl Human FamilyView
The Wellcome Trust Sanger Institute
Find Family ENSF00000000612 Lookup [e.g. ENSF00000000117] Help
Ensembl Protein Family ENSF00000000612 members
Consensus Annotation: ZYXIN.
Refresh
1 Mb
2 Mb
117.10 Mb
117.20
Ensembl Human MartView
The Wellcome Trust Sanger Institute
Home Human What's New BLAST SSAHA EnsMart Export Data Download Disease Browser Docs
Homology Matches
These gene(s) have been identified as put
Mus musculus ENSMUSG00000041301
Rattus norvegicus ENSRNOG00000008284
Fugu rubripes SINFRUG00000125655
Export Data
Export gene data in EMBL, GenBank or FAST
SNP legend

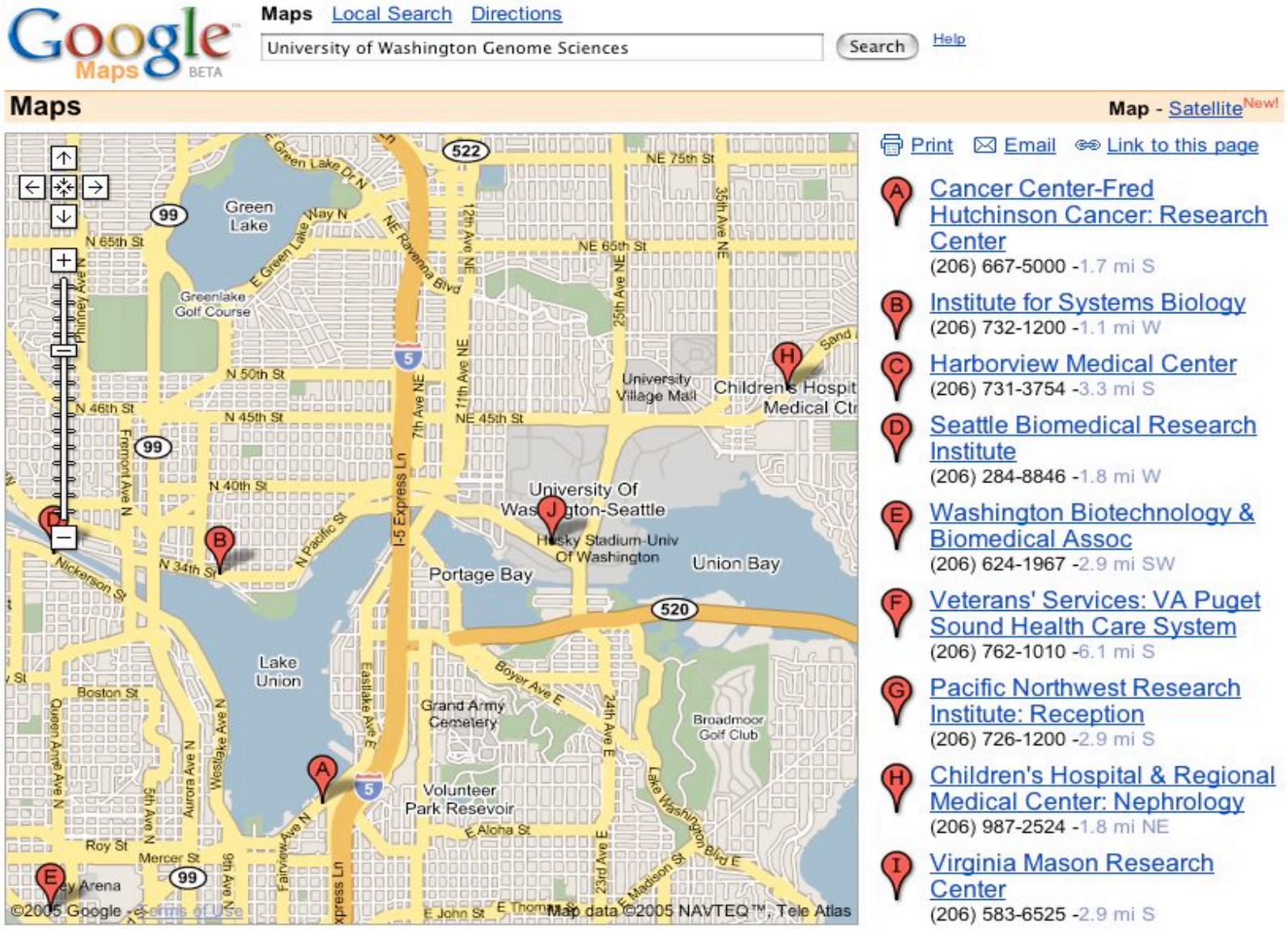
Under 10% data in Ensembl is sequence

Ensembl Gene ID	External Gene ID	Start Position (bp)	End Position (bp)	SWISSPROT I
ENSG00000137364	TPMT	18186685	18212843	TPMT_HUMAN
ENSG00000124795	DEK	18280050	18320277	DEK_HUMAN
ENSG00000010704	NM_139010	26130307	26139915	
ENSG00000010704	NM_139010	26130307	26139915	
ENSG00000010704	NM_139010	26130307	26139915	
ENSG00000010704	NM_139010	26130307	26139915	
ENSG00000010704	NM_139010	26130307	26139915	
ENSG00000010704	NM_139010	26130307	26139915	
ENSG00000010704	NM_139010	26130307	26139915	
ENSG00000010704	NM_139010	26130307	26139915	
ENSG00000010704	NM_139010	26130307	26139915	HFE_HUMAN
ENSG00000010704	NM_139010	26130307	26139915	
ENSG00000111997	NEU1	31898823	31902526	NER1_HUMAN
ENSG00000111997	NEU1	31898823	31902526	NER1_HUMAN
ENSG00000166278	C2	31957539	31975237	CO2_HUMAN
ENSG00000166301	STK19	32001436	32011003	ST19_HUMAN
ENSG00000166308	C4A	32011614	32032238	CO4_HUMAN
ENSG00000166308	C4A	32011614	32032238	CO4_HUMAN
ENSG00000168482	Q96NU8	32035126	32038461	
ENSG00000168482	Q96NU8	32035126	32038461	CPS1_HUMAN
ENSG00000173434	HLA-DQB1	32668141	32673755	HB21_HUMAN

Modern day maps: topography…



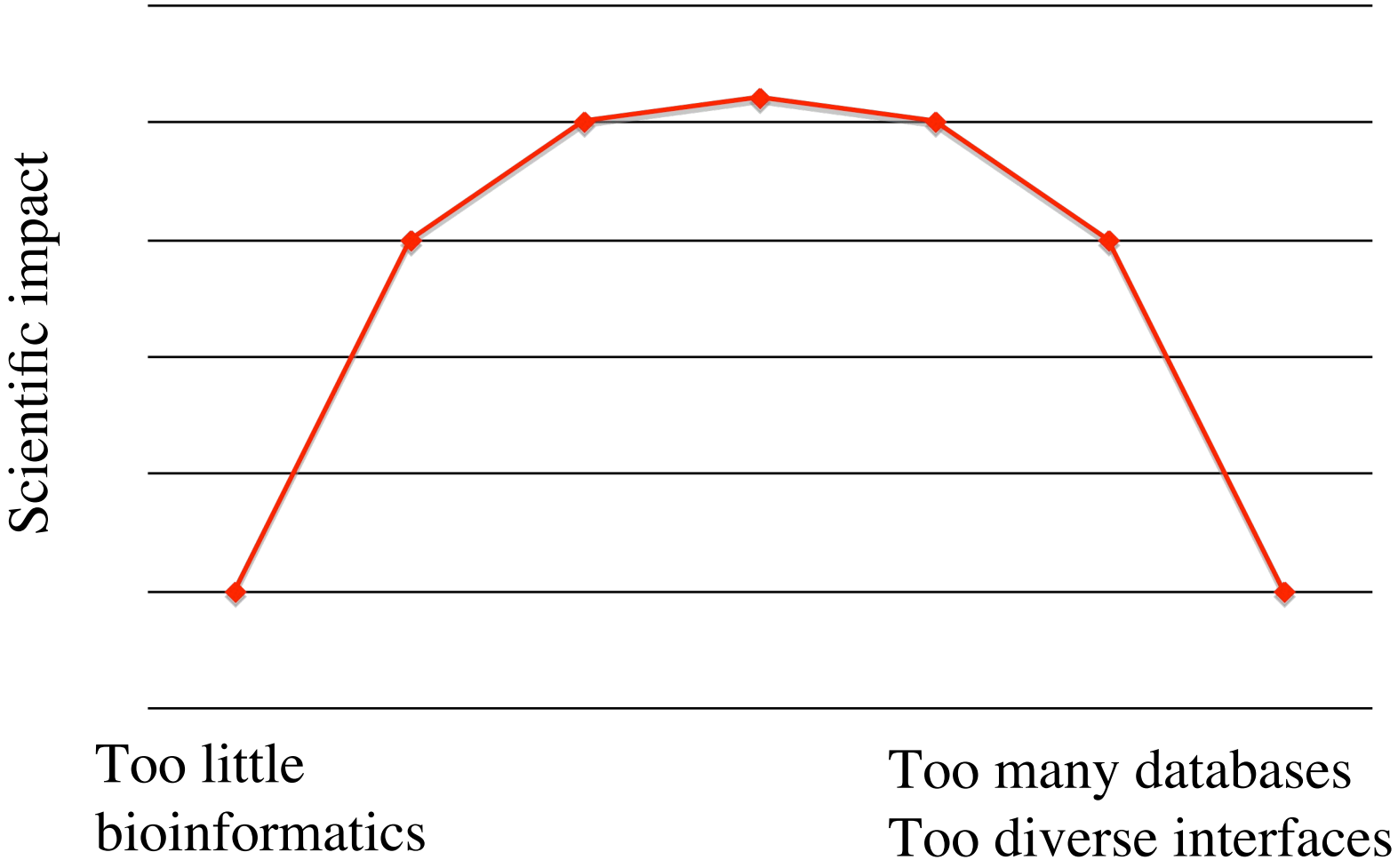
… plus annotation


Google Maps BETA
Maps Local Search Directions
University of Washington Genome Sciences
Search
Help
Maps
Map - Satellite New!
Print Email Link to this page
A Cancer Center-Fred Hutchinson Cancer: Research Center
(206) 667-5000 -1.7 mi S
B Institute for Systems Biology
(206) 732-1200 -1.1 mi W
C Harborview Medical Center
(206) 731-3754 -3.3 mi S
D Seattle Biomedical Research Institute
(206) 284-8846 -1.8 mi W
E Washington Biotechnology & Biomedical Assoc
(206) 624-1967 -2.9 mi SW
F Veterans' Services: VA Puget Sound Health Care System
(206) 762-1010 -6.1 mi S
G Pacific Northwest Research Institute: Reception
(206) 726-1200 -2.9 mi S
H Children's Hospital & Regional Medical Center: Nephrology
(206) 987-2524 -1.8 mi NE
I Virginia Mason Research Center
(206) 583-6525 -2.9 mi S
Green Lake
Greenlake Golf Course
University Of Was gton-Seattle
Husky Stadium-Univ Of Washington
Portage Bay
Union Bay
Lake Union
Grand Army Cemetery
Broadmoor Golf Club
Volunteer Park Resevoir
University Village Mall
Children's Hospit Medical Ctr
©2005 Google - Terms of Use
Map data ©2005 NAVTEQ™, Tele Atlas

Data integration

- Complete genomes provide the framework to pull all biological data together such that each piece says something about biology as a whole
- Biology is too complex for any organisation to have a monopoly of ideas or data
- The more organisations provide data or analysis separately, the harder it becomes for anyone to make use of the results

Utility of bioinformatics

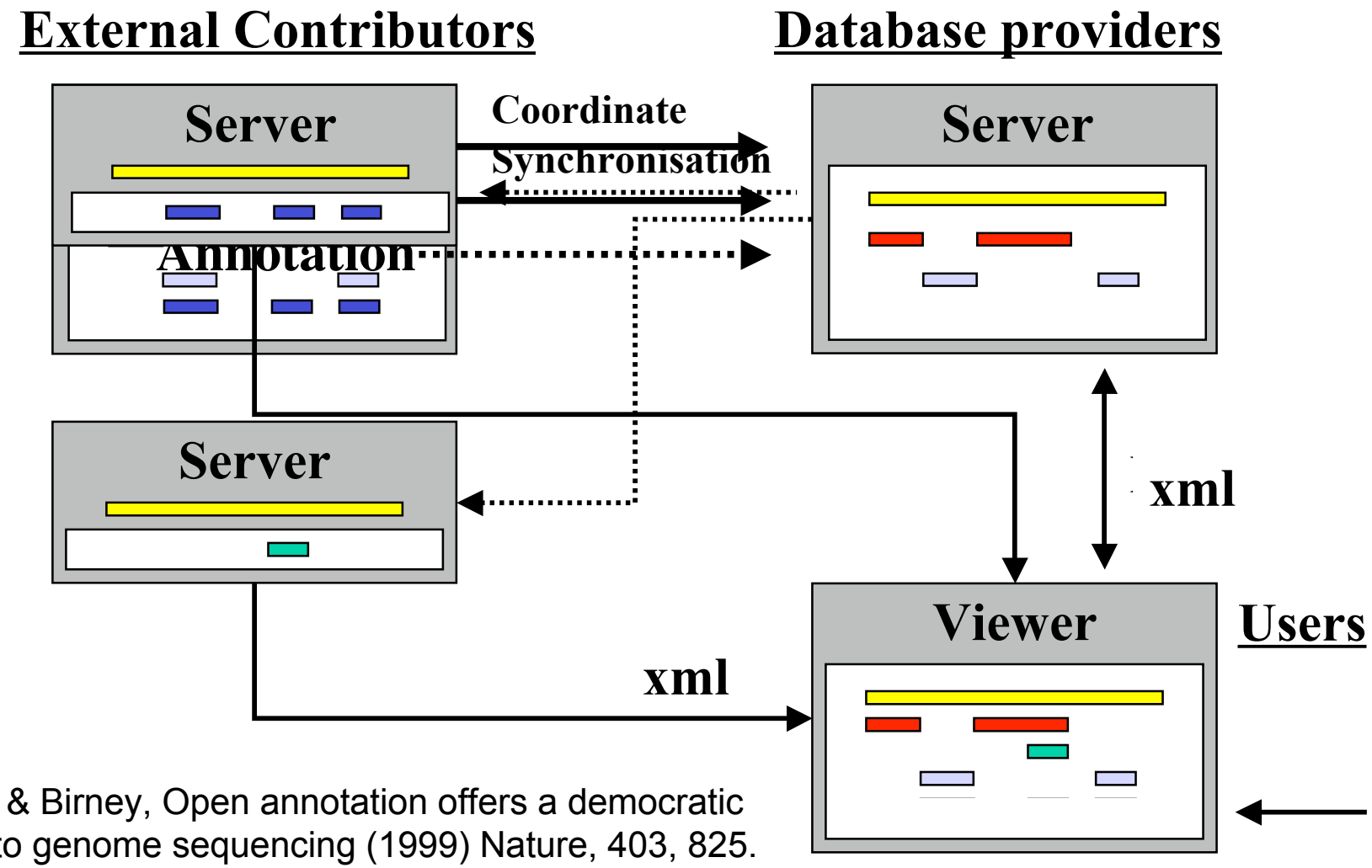


Split data and presentation

- Databases responsible for curating data and serving it as primitive datatypes defined by open standards (high cost)

- Different front ends or components of front ends compete for users (development of each low cost) c.f. browsers.

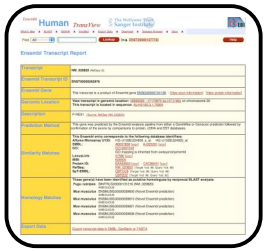
Distributed Annotation



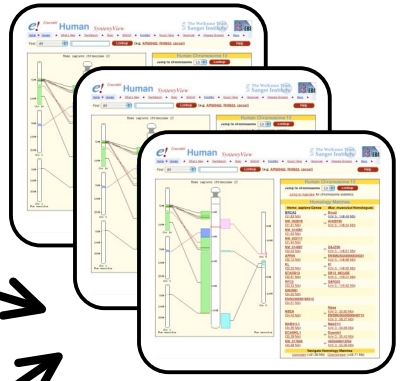
Hubbard & Birney, Open annotation offers a democratic solution to genome sequencing (1999) Nature, 403, 825.
See: http://biodas.org/ Original proposal: Lincoln Stein et al

DAS v Web

Web Model:

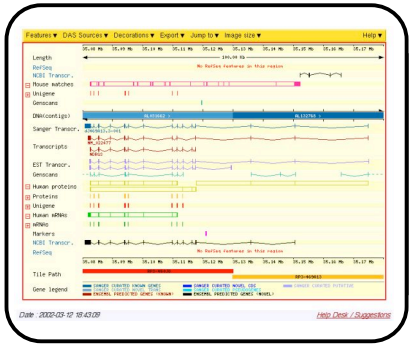


links

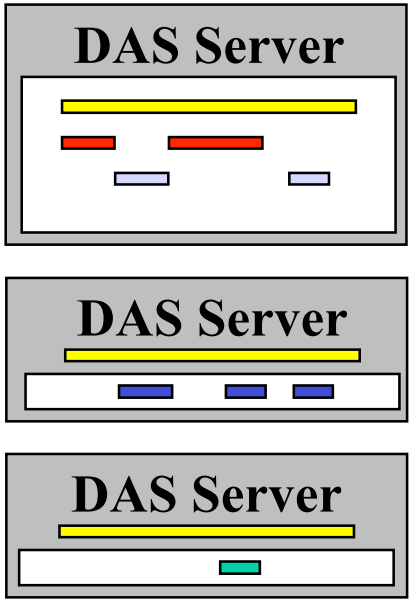


Different Web sites
Different interfaces
No integration

DAS Model:



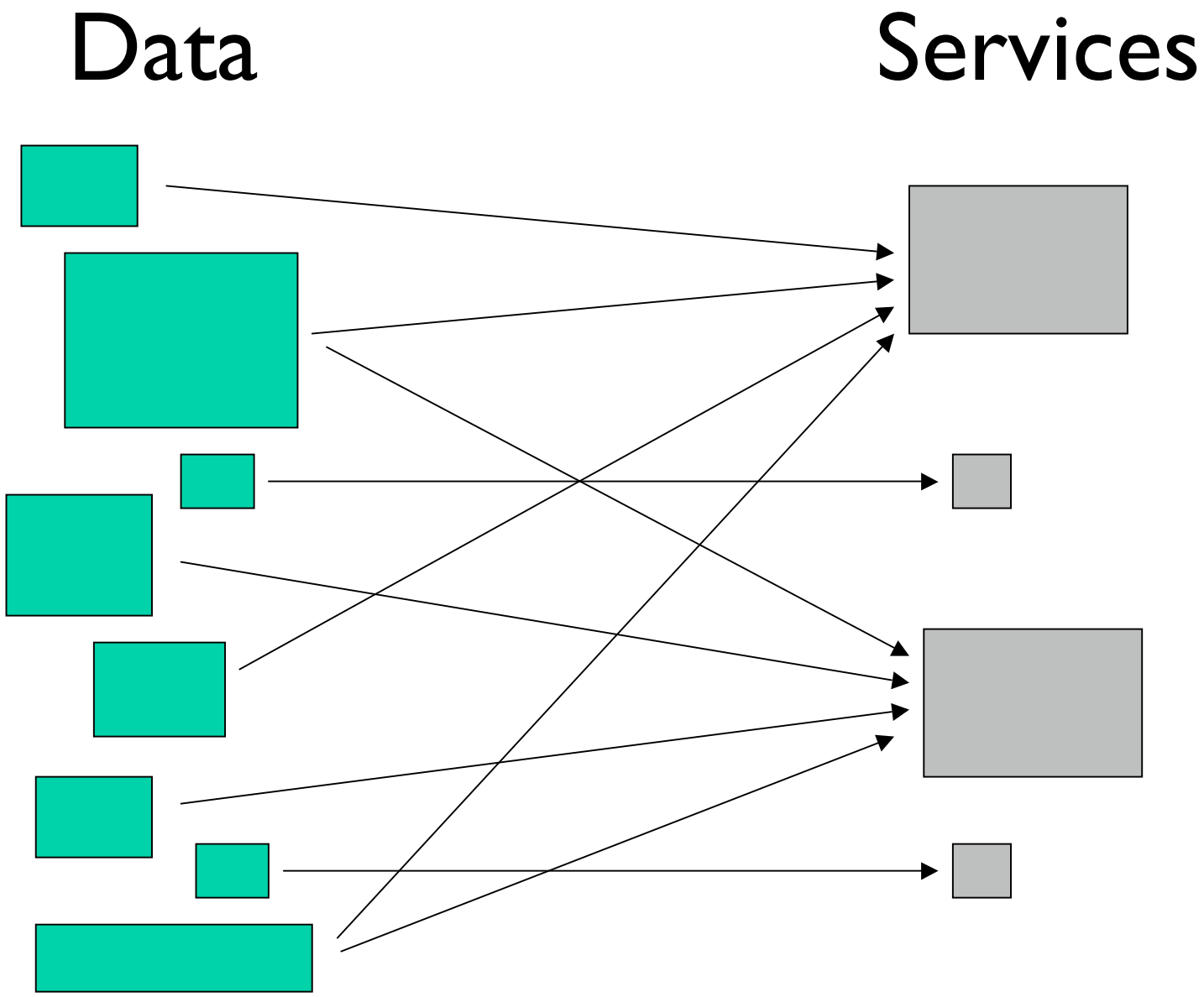
Different DAS sites
Automatic Integration
Single interface

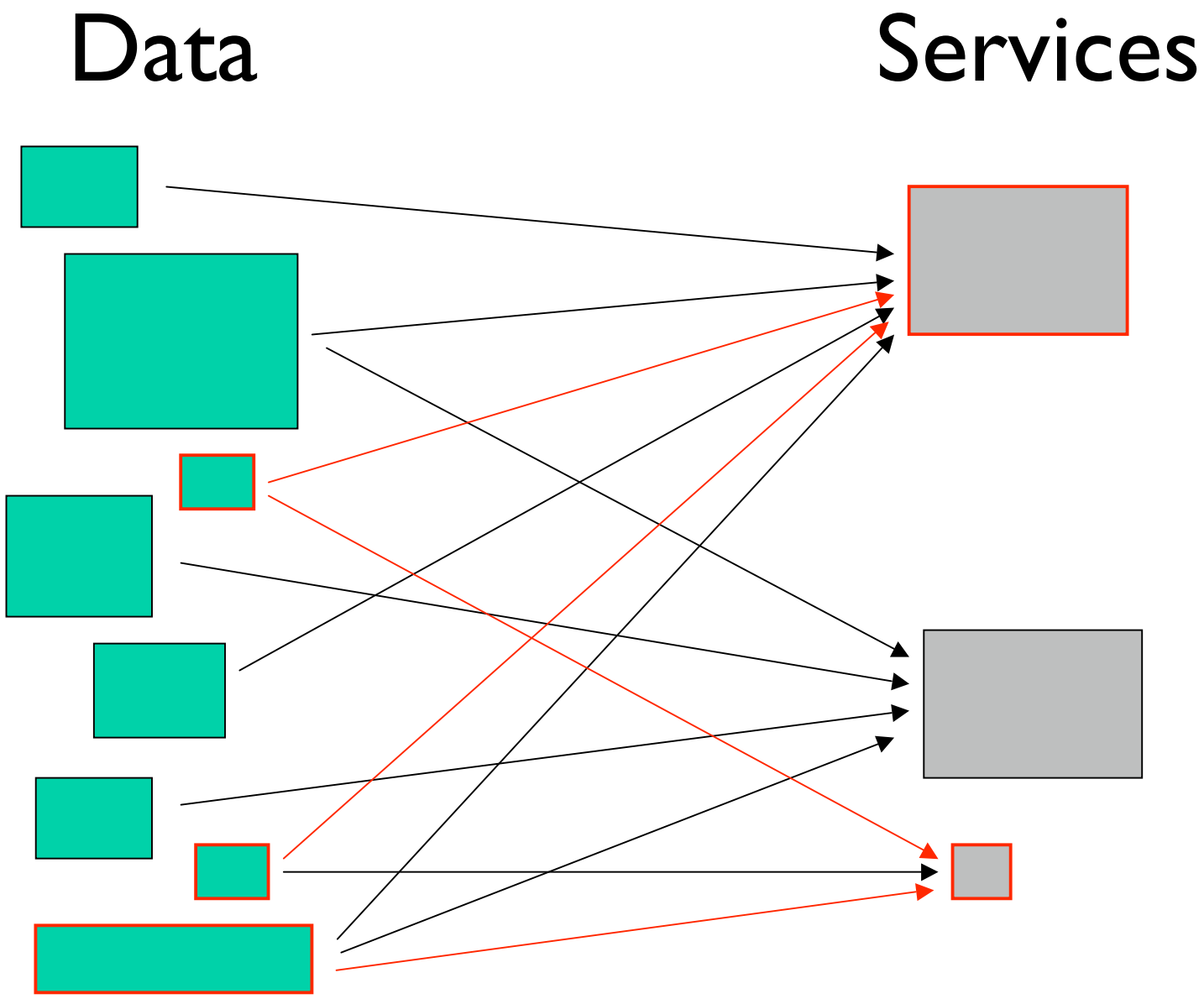


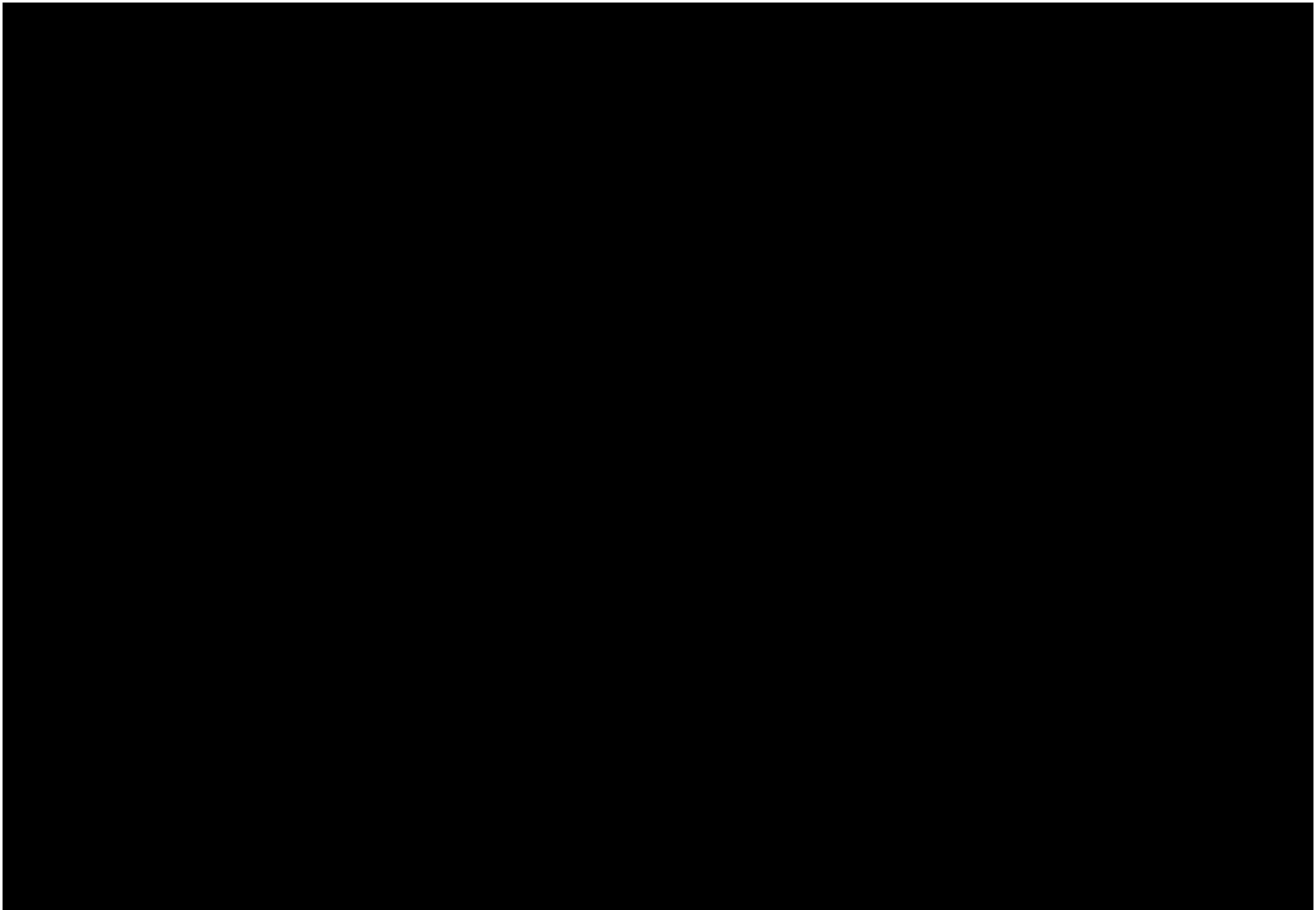
DAS in a nutshell

- Standardized set of web services
  - Reference servers (the sequence)
  - Annotation servers (features: chr:start-end)

  - Alignment servers (chr:start-end matches chr:start-end)
  - Identifier based servers (ref item X rather than coordinate)


Data
Services

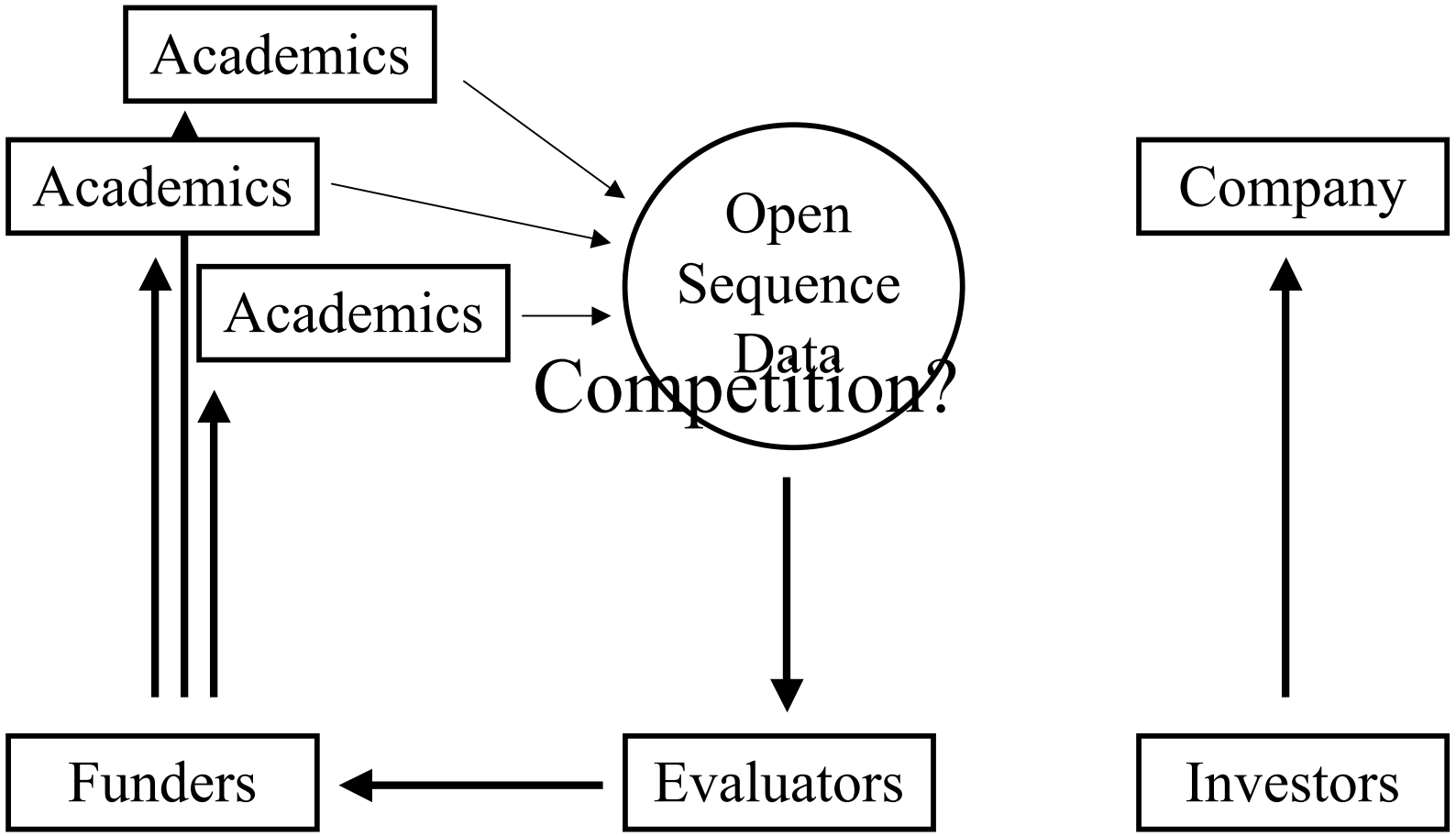

Data
Services



What is the problem with government?

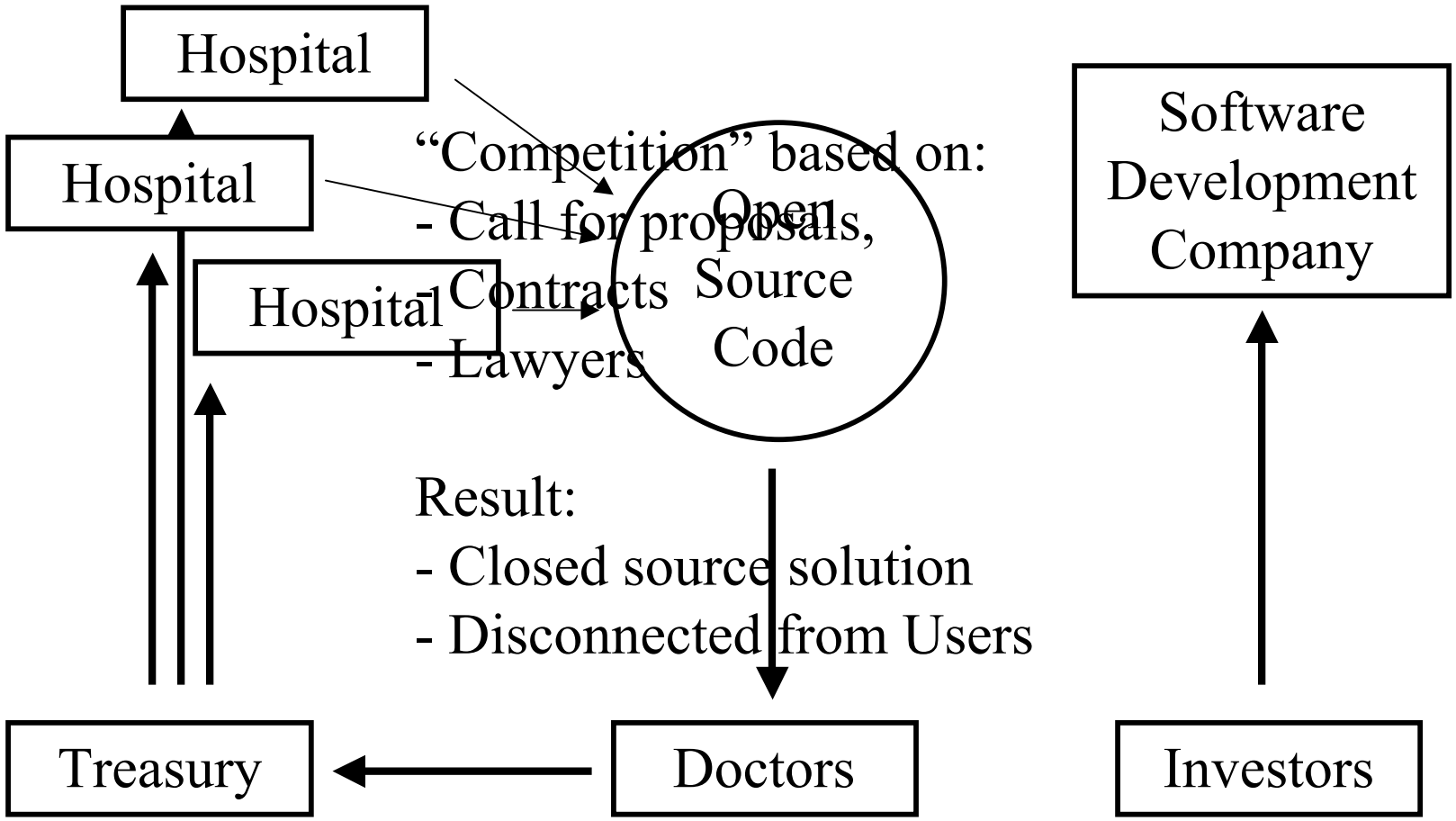
- Where are open ideas are changing policy? Why?
- +ve Switching publishing to Open Access publishing model
  - Less risky than most since self sustaining
    - Value protected through peer review
- -ve Long term support for resources that cost money
  - e.g Science infrastructure
    - Continuously expanding; mechanisms for merging, eliminating ones with little value poor, or perceived to be poor
- Genuine worries about whether change will affect stability
  - Potential unknown pernicious incentives
  - If you change the system from one that currently partly works to one that breaks, we get blamed
- Lack of technically aware people in government?
  - Culture of contracting out and lawyers (self reinforcing)

The Human Genome Project



Weekly conference calls
Competitive Collaboration

Public computer projects



Continuous evaluation by Doctors
Competitive Collaboration Innovation

Market properties of openness

- "Optimal allocation of resources based on transparency of information"
- Evaluation can be *continuous* and *in parallel* with research
  - Researchers can stay more focused on their work, rather than writing heavy grants and reports
  - Potential innovation and competition in evaluation as data is open to all

Wellcome Trust commissioned two reports on Open Access publishing

September 2003

"An Economic Analysis of Scientific Research Publishing"

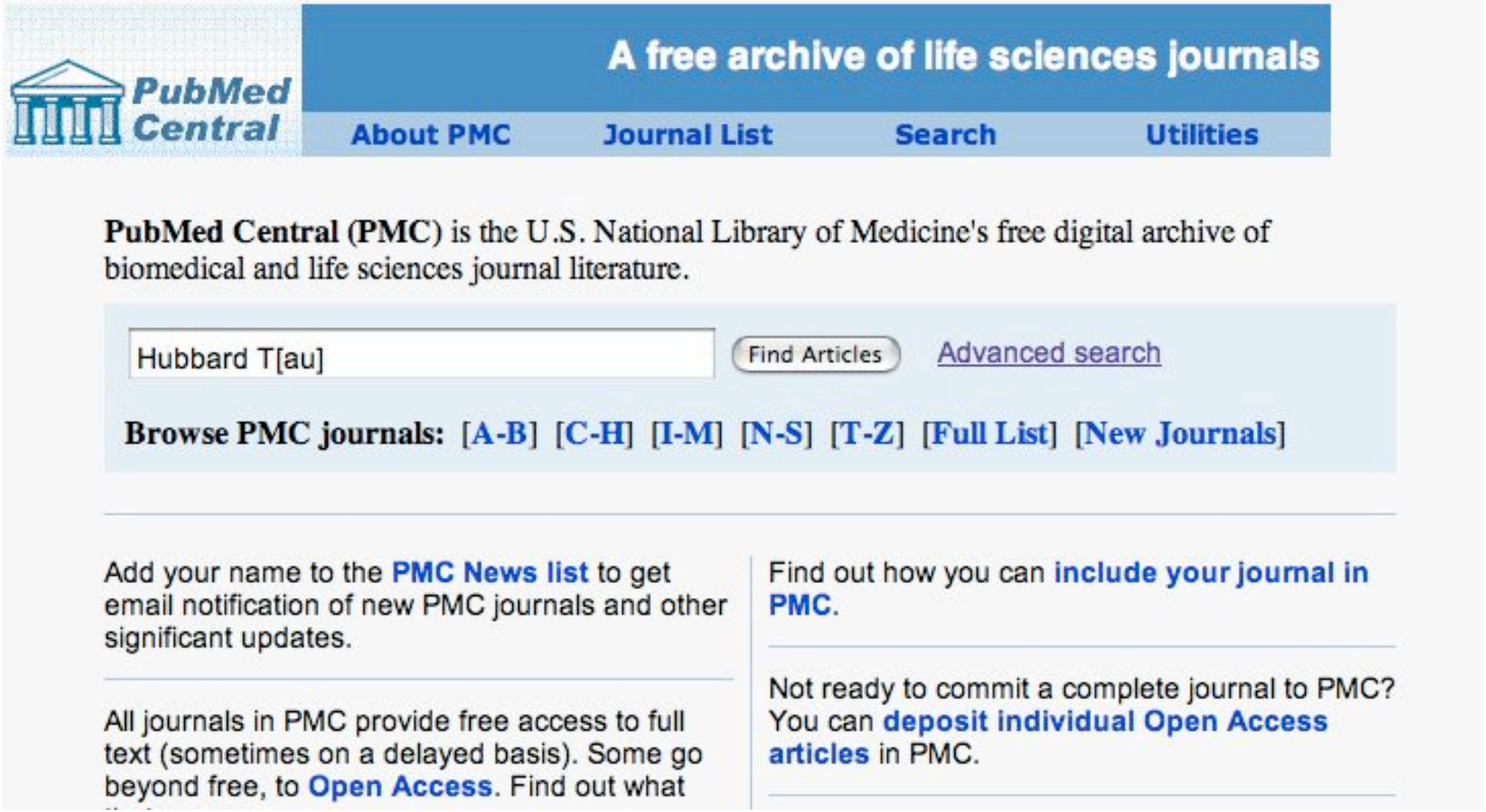
April 2004

"Costs and Business Models in Scientific Research Publishing"

Conclusions: Open access could cut scientific publishing costs by 30 per cent and still provide a viable business model

UK Parliamentary select committee is now investigating the issue of open access

Open access publishing allows anyone to analysis and organise scientific text


PubMed Central
A free archive of life sciences journals
About PMC
Journal List
Search
Utilities
PubMed Central (PMC) is the U.S. National Library of Medicine's free digital archive of biomedical and life sciences journal literature.
Hubbard T[au]
Find Articles
Advanced search
Browse PMC journals: [A-B] [C-H] [I-M] [N-S] [T-Z] [Full List] [New Journals]
Add your name to the PMC News list to get email notification of new PMC journals and other significant updates.
All journals in PMC provide free access to full text (sometimes on a delayed basis). Some go beyond free, to Open Access. Find out what
Find out how you can include your journal in PMC.
Not ready to commit a complete journal to PMC? You can deposit individual Open Access articles in PMC.

Openness makes assessment easier

- Researchers frequently assessed on where papers are published
  - Nature, Science have some of high impact factors
  - Crude measure: many papers published in top journals never cited
- Open access publication has potential to provide granting agencies with much more evaluation of research output

New context of data sharing

- Internet facilitates easy data sharing
  - Previously practical issues as big as anything else
- Its our data, our research: we paid for it!
  - Both funders and public beginning to realise what they have been missing
- Complex problems require open models
  - Some problems will only be solved by collaborative analysis of shared data collections

Open approaches only solution for ultra-complex problems?

- Benefits from genome so far limited since there is a lot to understand
- Microsoft experience with Vista suggests limit on centralized model for complex systems
- Science is already highly cooperative, however perhaps the genome will only be successfully exploited in a distributed open data environment

“open” verses “closed”

- Biological data: Public/Private
- Scientific Databases: Public/Private
- Literature: Open Access/Commercial Journals
- Software: Open Source (GNU/LINUX)/Closed
- Heathcare R&D: Public based/Patent based

- Research is driven by openness, Internet
- Business models can be based on it (e.g. IBM linux)

open source goes mainstream…

Economist.com **OPINION**

Open source

Beyond capitalism?

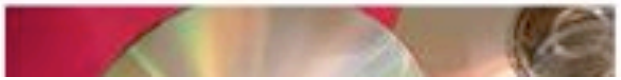
Jun 10th 2004
From The Economist print edition

The open-source model can be applied to goods other than software, but it has its limits

THAT "open source" is a good way to make software is beyond question. For those unfamiliar with the term, the open-source model allows many people to collaborate on the development of a piece of software by making its underlying programming instructions, or source code, open to everyone, usually by publishing them on the internet. The resulting program is then given away too: open-source software is shared, not sold. Commercial software vendors, by contrast, jealously guard their source code because only by keeping it secret can they protect their ability to demand money for their products.

By far the best-known example of open-source software is Linux, an operating system that is maintained by volunteers around the world, runs on everything from wristwatches to mainframes and

Alamy



…and extends to medicines

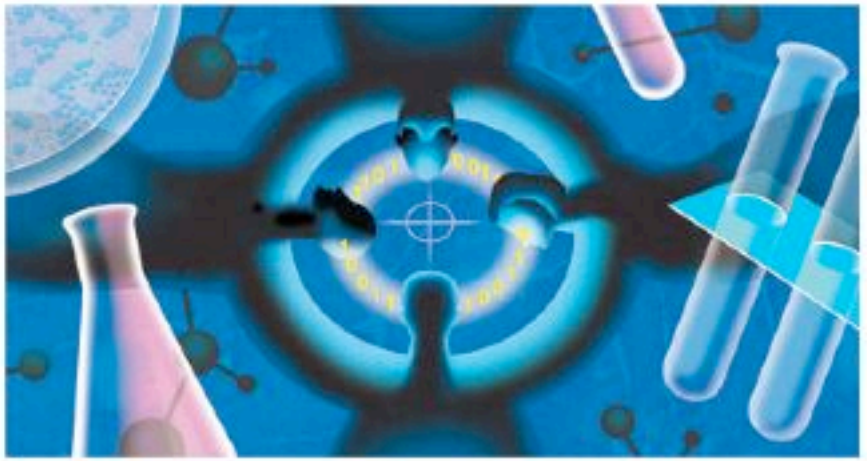
Economist.com

SCIENCE
TECHNOLOGY QUARTERLY

REPORTS

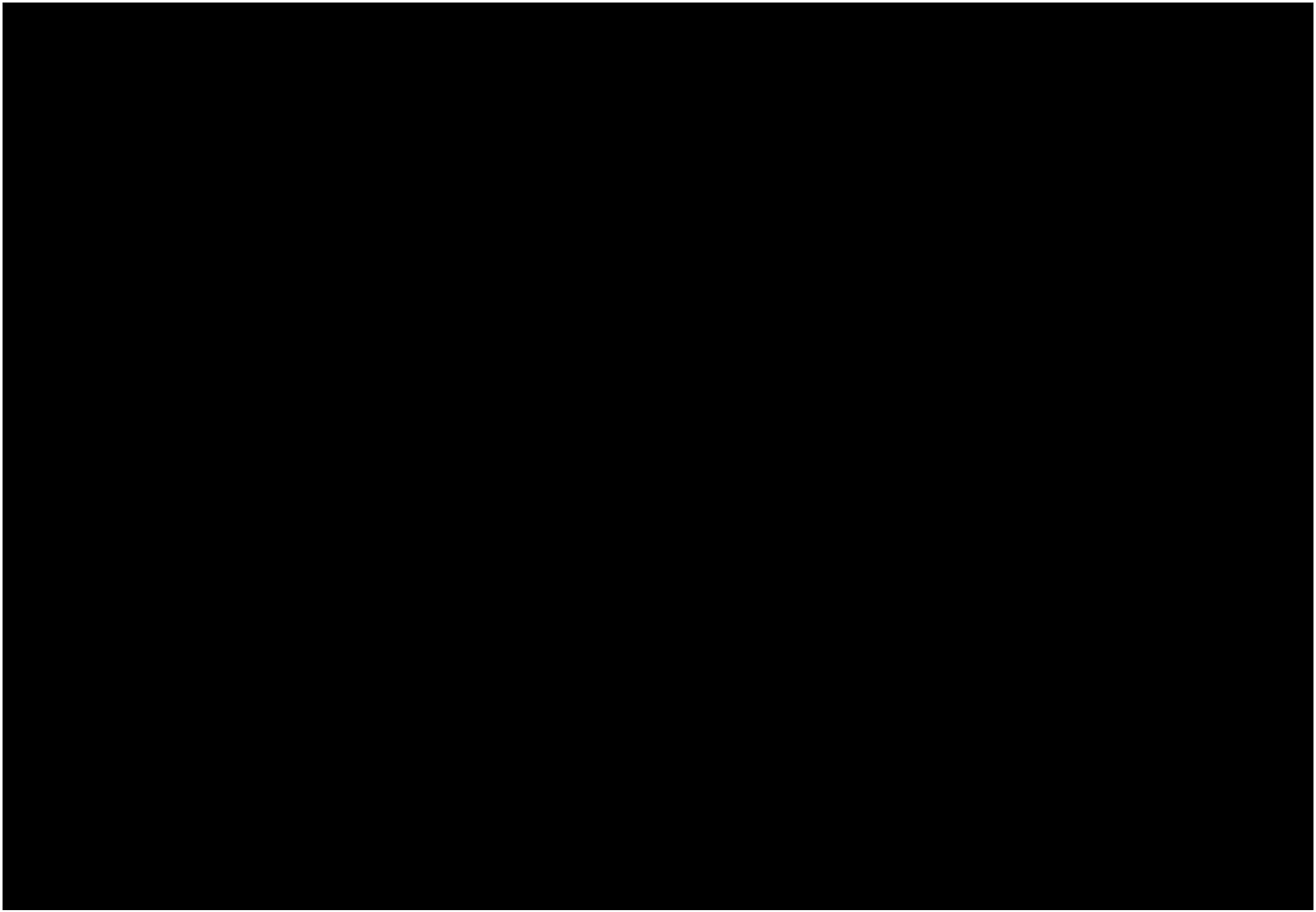
An open-source shot in the arm?

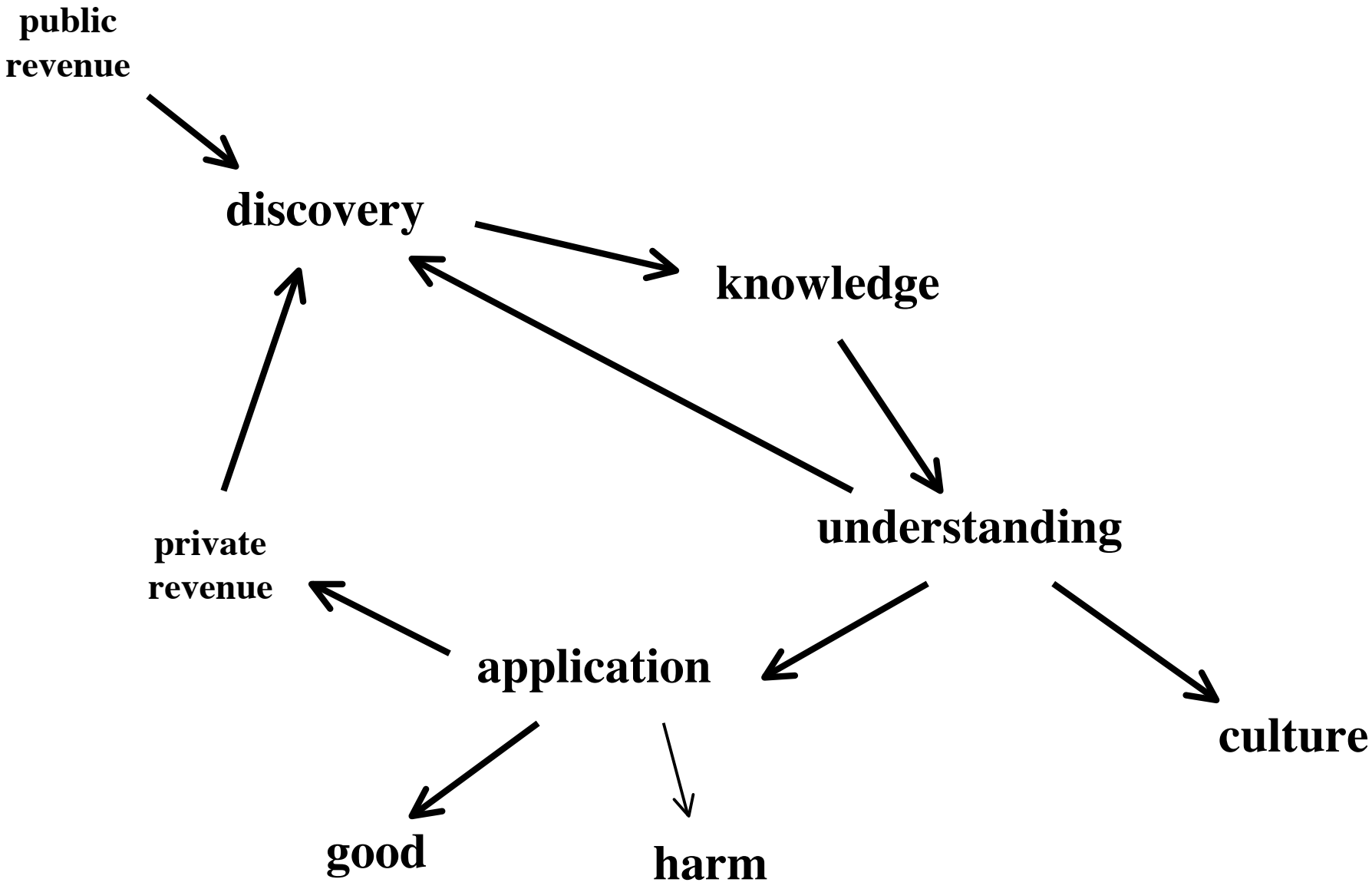
Jun 10th 2004
From The Economist print edition

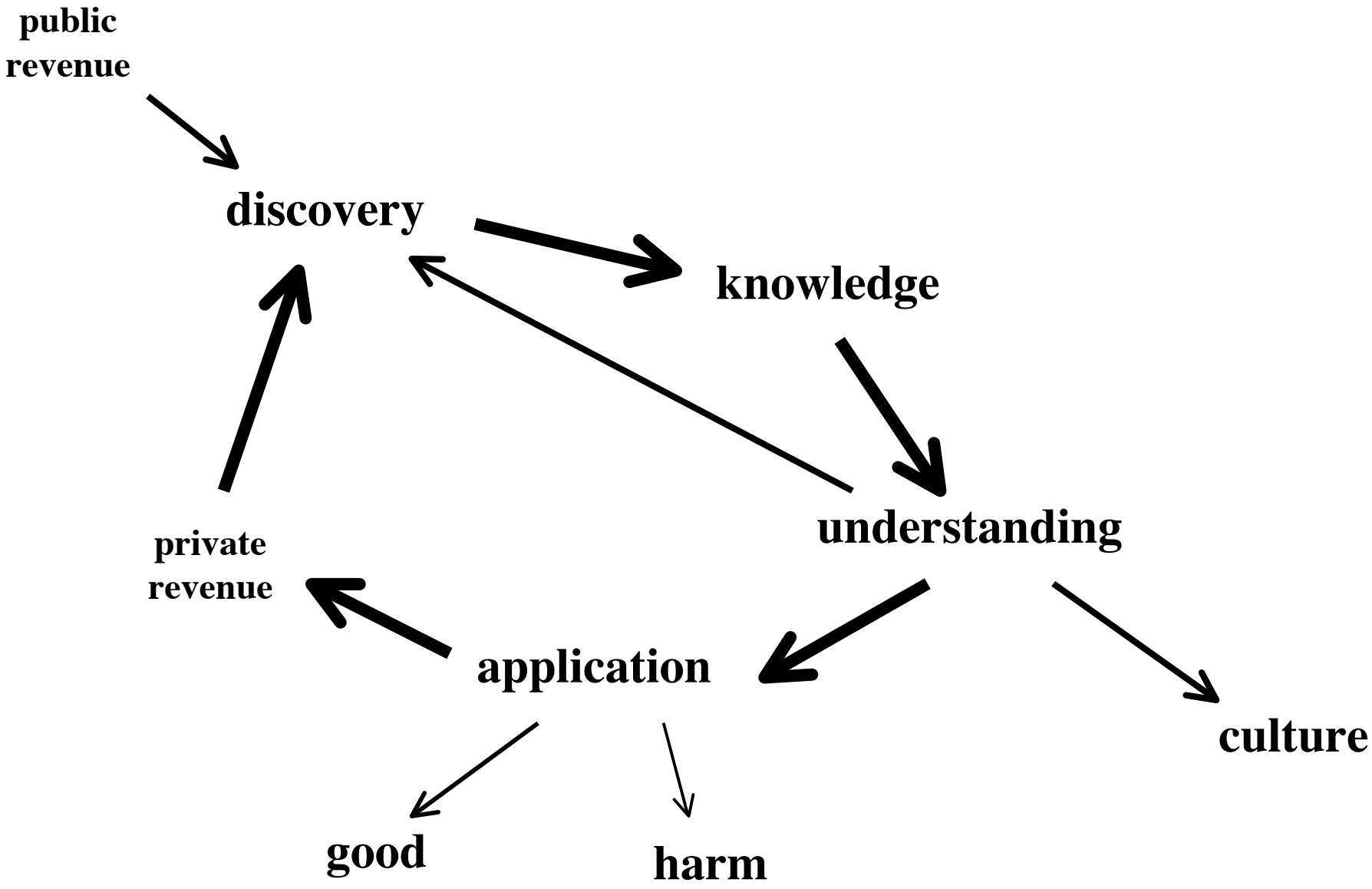


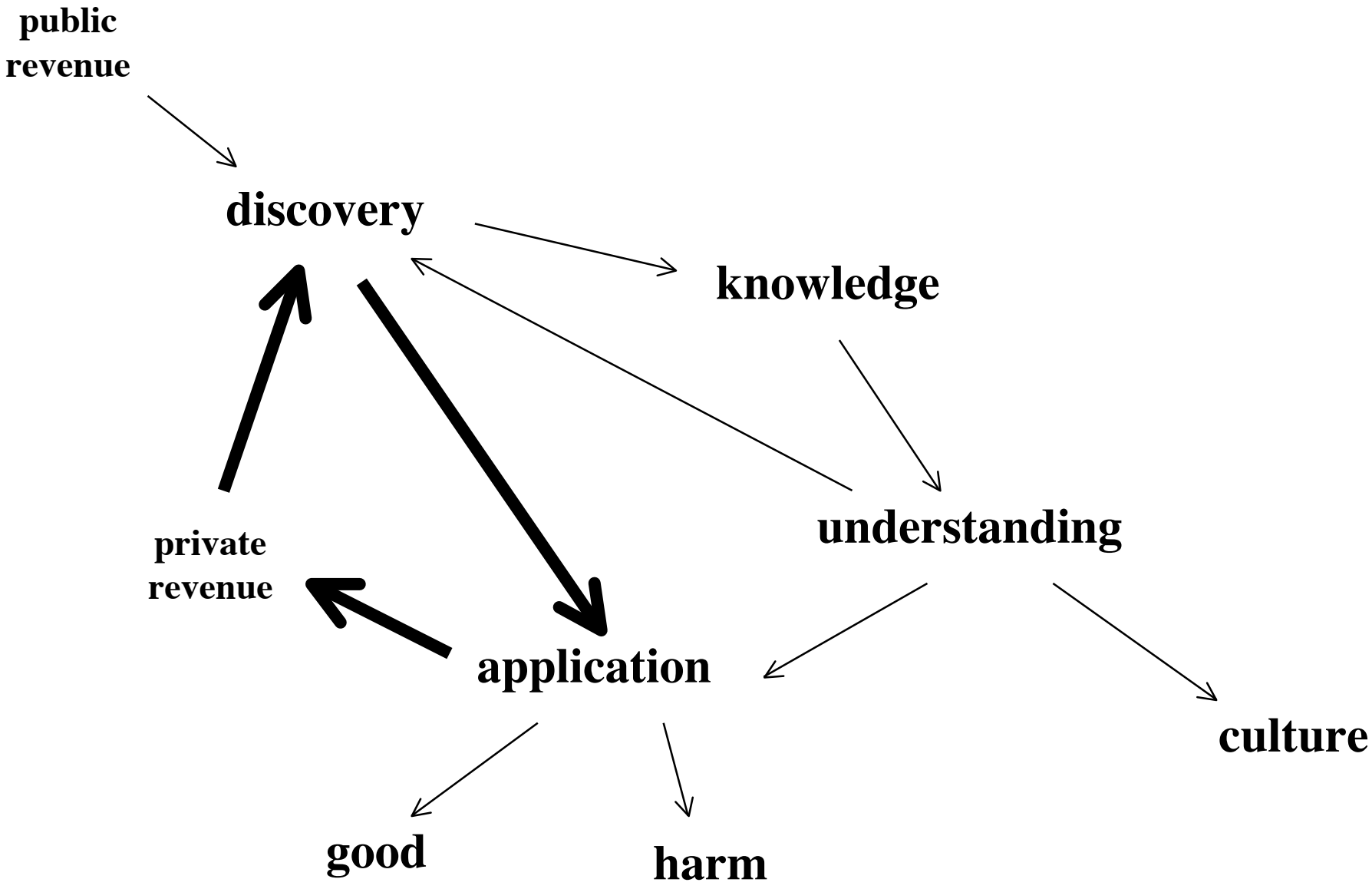
Medicine: The open-source model is a good way to produce software, as the example of Linux shows. Could the same collaborative approach now revitalise medical research too?

CAN goodwill, aggregated over the internet, produce good medicine? The current approach to drug discovery works up to a point, but it is far from perfect. It is costly to develop medicines and get regulatory approval. The patent system can foreclose new uses or enhancements by outside researchers. And there has to be a consumer willing (or able) to pay for the resulting drugs, in order to justify the cost of drug development. Pharmaceutical companies have little incentive to develop treatments for diseases that particularly afflict the poor, for example, since the people who need such treatments most may not be able to afford them.

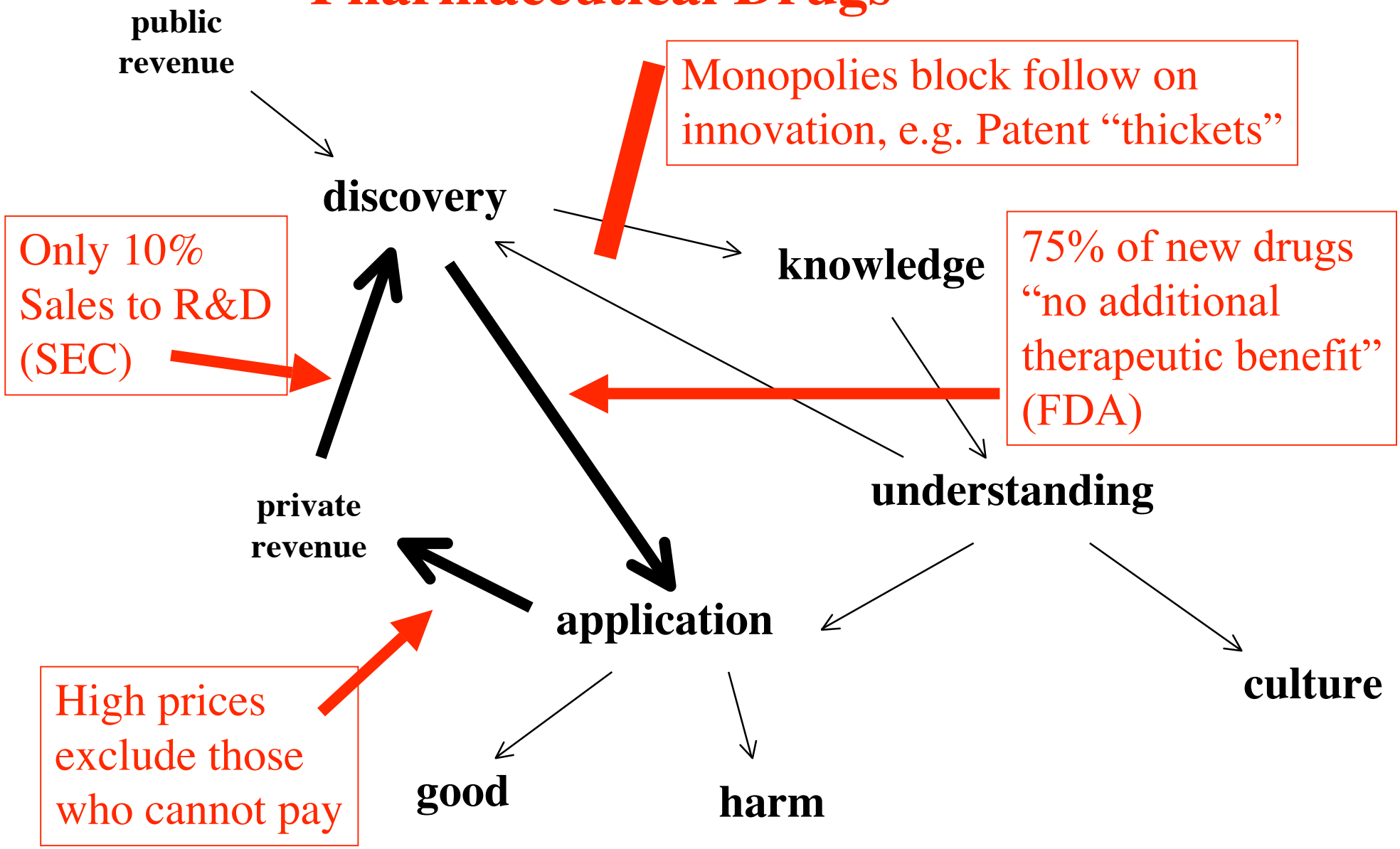



public
revenue
discovery
knowledge
understanding
private
revenue
application
culture
good
harm


public
revenue
discovery
knowledge
understanding
private
revenue
application
culture
good
harm


public
revenue
discovery
knowledge
understanding
private
revenue
application
culture
good
harm

Pharmaceutical Drugs



Scale of problem

- Every day over 35,000 people die from infectious diseases such as AIDS, malaria tuberculosis, and other neglected diseases such as Leishmaniasis, Chagas disease and sleeping sickness.
- Where drugs do exist, millions are unable to access them, frequently due to high prices.

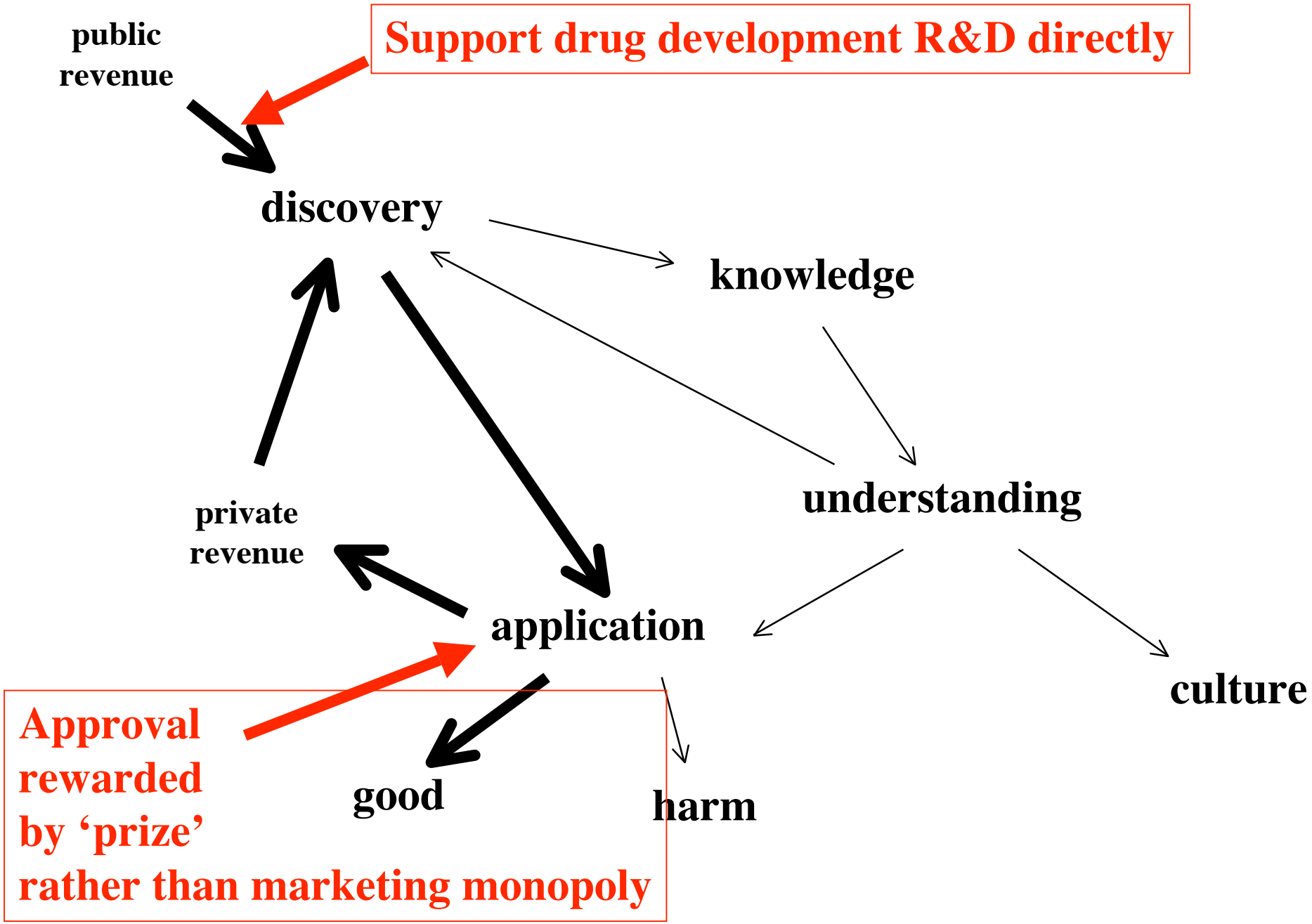
Health Comparisons

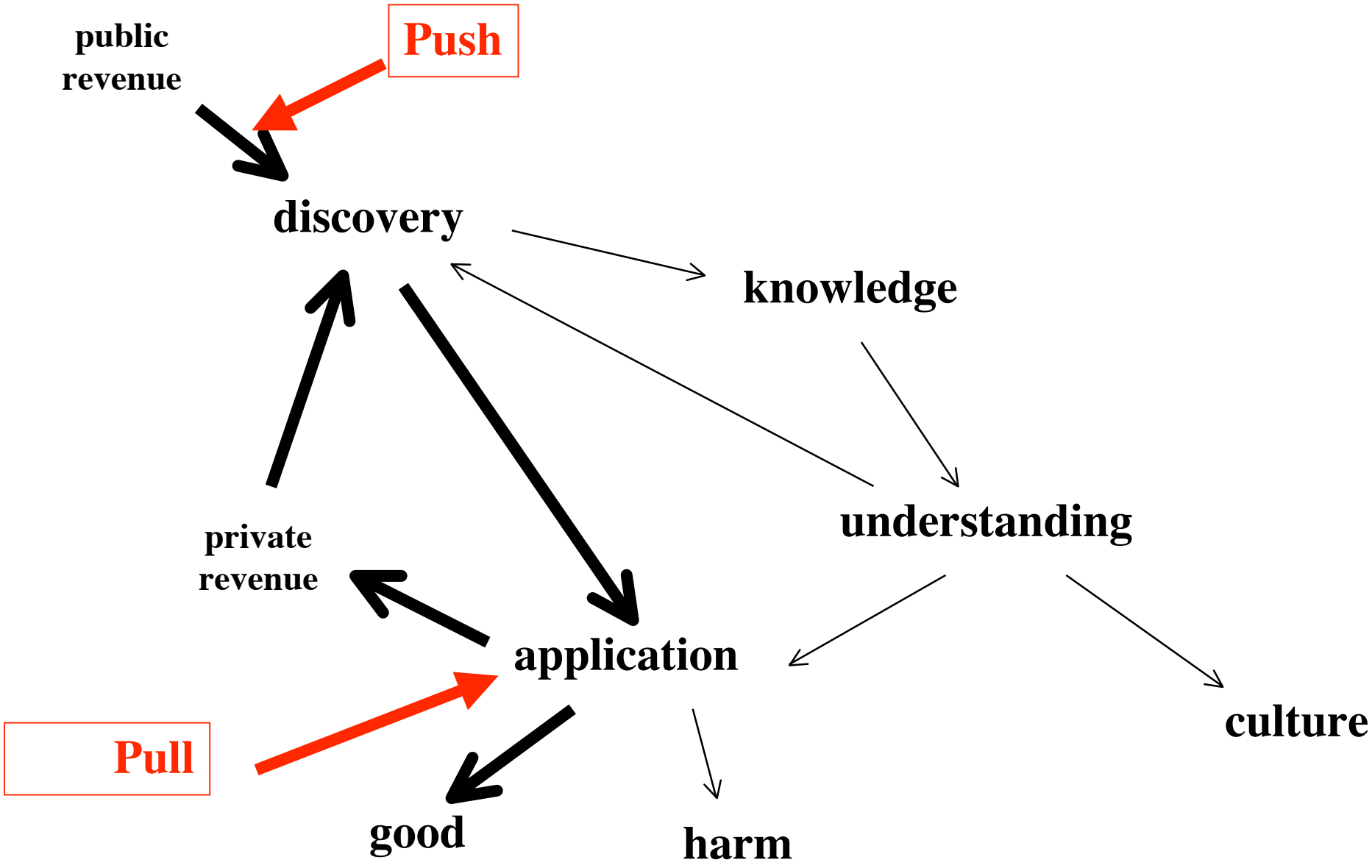
Country	Annual $ health spend per person	Health spend % GDP	~equivalence of UK prescription
USA	5,491	15.2	
UK	2,110	8.0	£6.65
Thailand	68	3.3	£80.00
India	24	4.8	£350.00
Malawi	16	9.3	£1000.00

Two problems

- Innovation:
  how to support R&D for diseases on basis of need rather than market

- Access:
  how to provide access to medicines at affordable prices while still providing financial incentives for R&D


public revenue
Support drug development R&D directly
discovery
knowledge
understanding
private revenue
application
culture
Approval rewarded by 'prize' rather than marketing monopoly
good
harm


public
revenue
Push
discovery
knowledge
understanding
private
revenue
application
Pull
good
harm
culture

Some of the advantages

- Investment in R&D will follow criteria for prize fund awards: set to match health priorities
- Can incentivise development before market exists, e.g. antibiotics, vaccines
- Can reward primary and secondary innovators; allowing an immediate switch to best treatment
- Removes incentive to market excessively
- Implementation under a Treaty framework requires no central fund, no central planning

Recent progress in real world

- Global fund has spent $7 billion in 136 countries
- UNITAID: international air-ticket solidarity levy (founded 16/9/06) - will collect €300 million in 2007
- Advance Market Commitment (AMC) for vaccines launched, $1.5 billion (launched 9/2/07)

- Intergovernmental working group (IGWG) set up by WHO following CIPIH report (5/06), meeting now

- Campaigning NGOs (MSF, Oxfam, HAI, TWN, CPtech) played a key role putting this on the agenda
- Our involvement: http://www.whoscientistsletter.org/

ORIGINALITY
IS UNDETECTED
PLAGIARISM